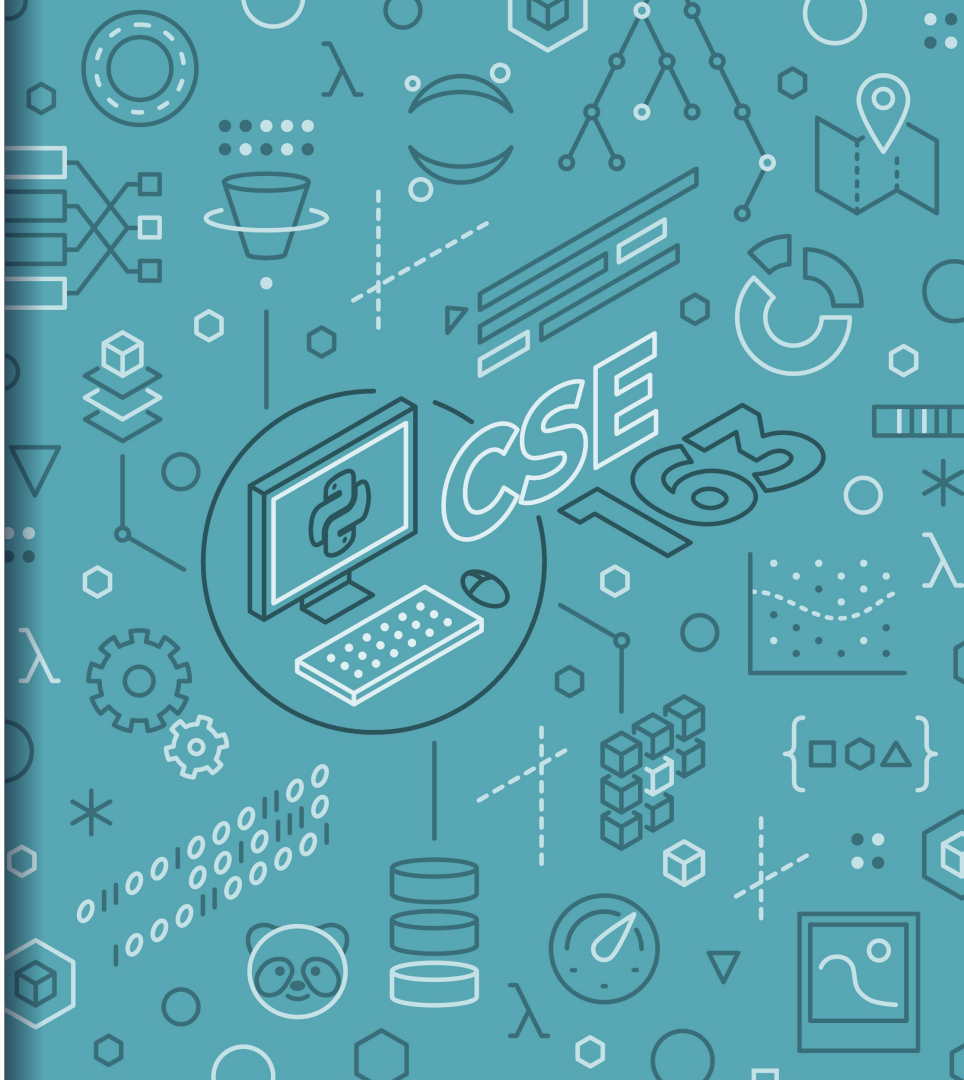


CSE 163

Missing Data & Time Series

Hunter Schafer

DataFrame

- One of the basic data types from pandas is a DataFrame
  - It's essentially a table with column and rows!

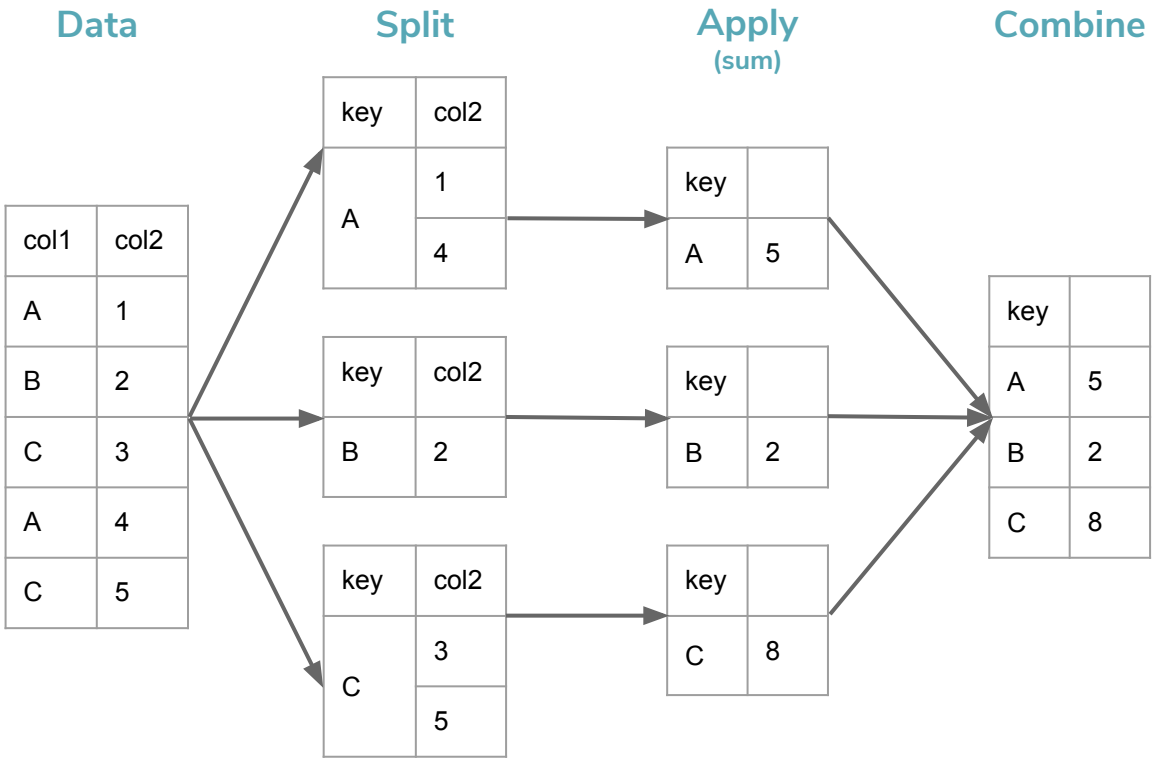
Columns

Index (row)

	id	year	month	day	latitude	longitude	name	magnitude
0	nc72666881	2016	7	27	37.672333	-121.619000	California	1.43
1	us20006i0y	2016	7	27	21.514600	94.572100	Burma	4.90
2	nc72666891	2016	7	27	37.576500	-118.859167	California	0.06

Group By

```
data.groupby('col1')['col2'].sum()
```



This Week

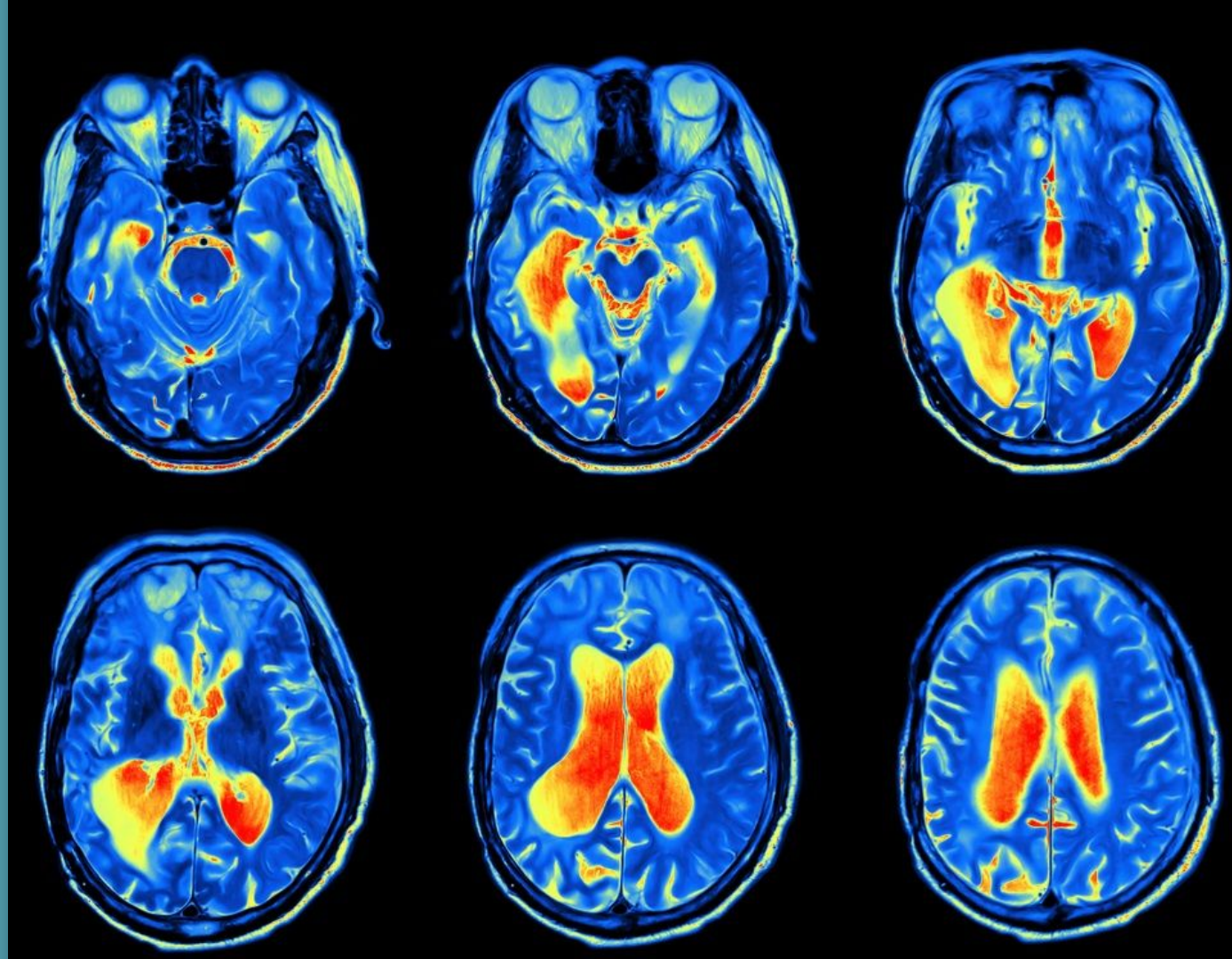
Data Science Libraries

- Monday
  - Missing Data
  - Time Series
  - Library: pandas
- Wednesday
  - Data Visualization
  - Library: seaborn
- Friday
  - Machine Learning
  - Library: scikit-learn

What to Learn

- This week, we are learning more about pandas and learning 2 new libraries
- Memorizing the function calls and parameters is ridiculous
  - No one memorizes this stuff!
  - This is what documentation is for!
- Much more important to understand the big ideas behind what the library call is doing
  - You might use a different library in the future
  - They change from version to version
- Don't try to write every bit of syntax down, focus on the big ideas behind what we are trying to solve and use the slides and lecture notes as a resource.
- On the exam, we will provide shortened documentation so you don't have to memorize the method calls

fMRI



Missing Data

- Most data in the world is messy and not in a form you want
  - Most common: Missing data
- Pandas uses “Not a Number” (NaN) to represent missing data
- Most times, it will just ignore them in computations but NaN can be a common source of bugs!
- Useful pandas functions

Detecting for missing data
`isnull()`
`notnull()`
Changing/Removing missing data
`dropna()`
`fillna()`

Demo

Sorting

- Sorting your data is a very common task
  - Either for presentation or finding the top-k
- Very easy in pandas!
  - Note: All of these return new `DataFrames`

```
# Sort data
data.sort_values('column')
data.sort_index()
# Find top-k
data.nlargest(10, 'column')
```

Demo

Keyword Arguments

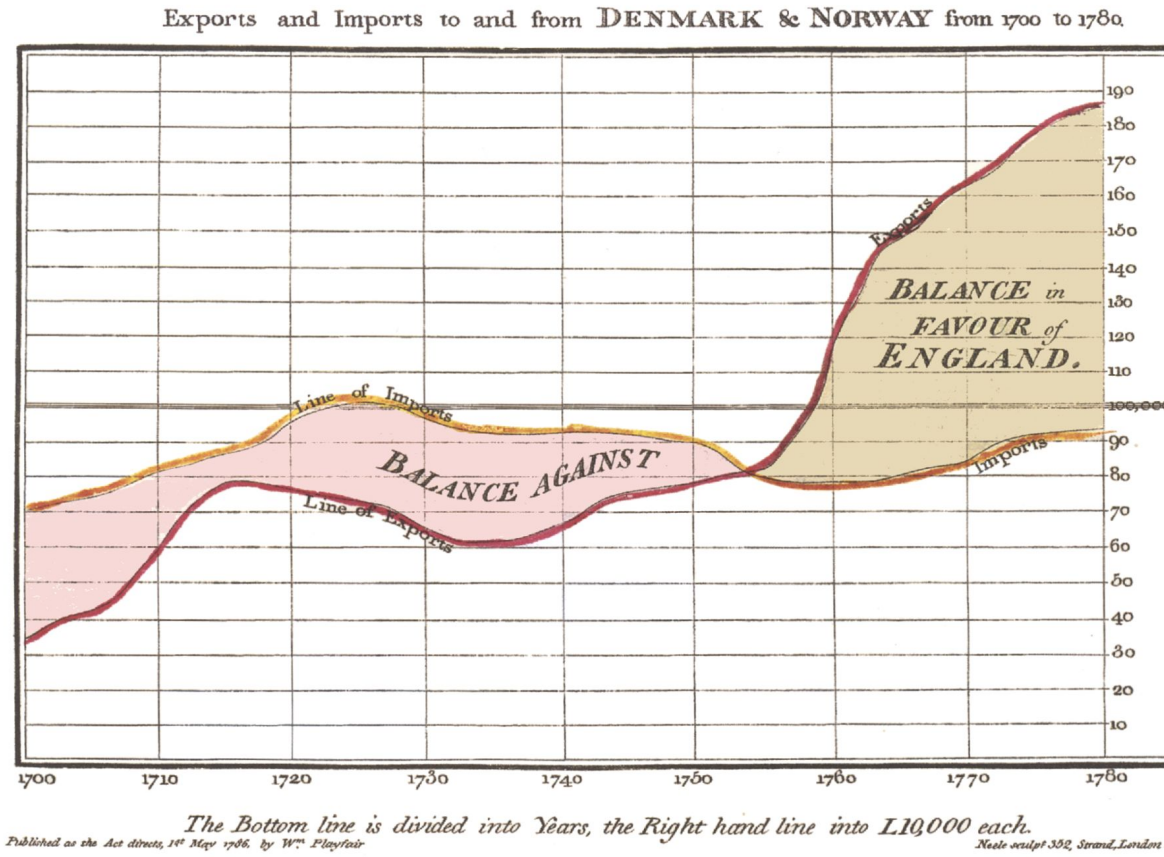
```
def div(a, b):
    return a / b

div(2, 3)
```

- How does Python know that a is 2 and b is 3?
  - Arguments are determined by position
- Python also allows you to pass by name instead
  - Library calls usually take MANY arguments (with defaults), much more convenient to specify the ones you by name

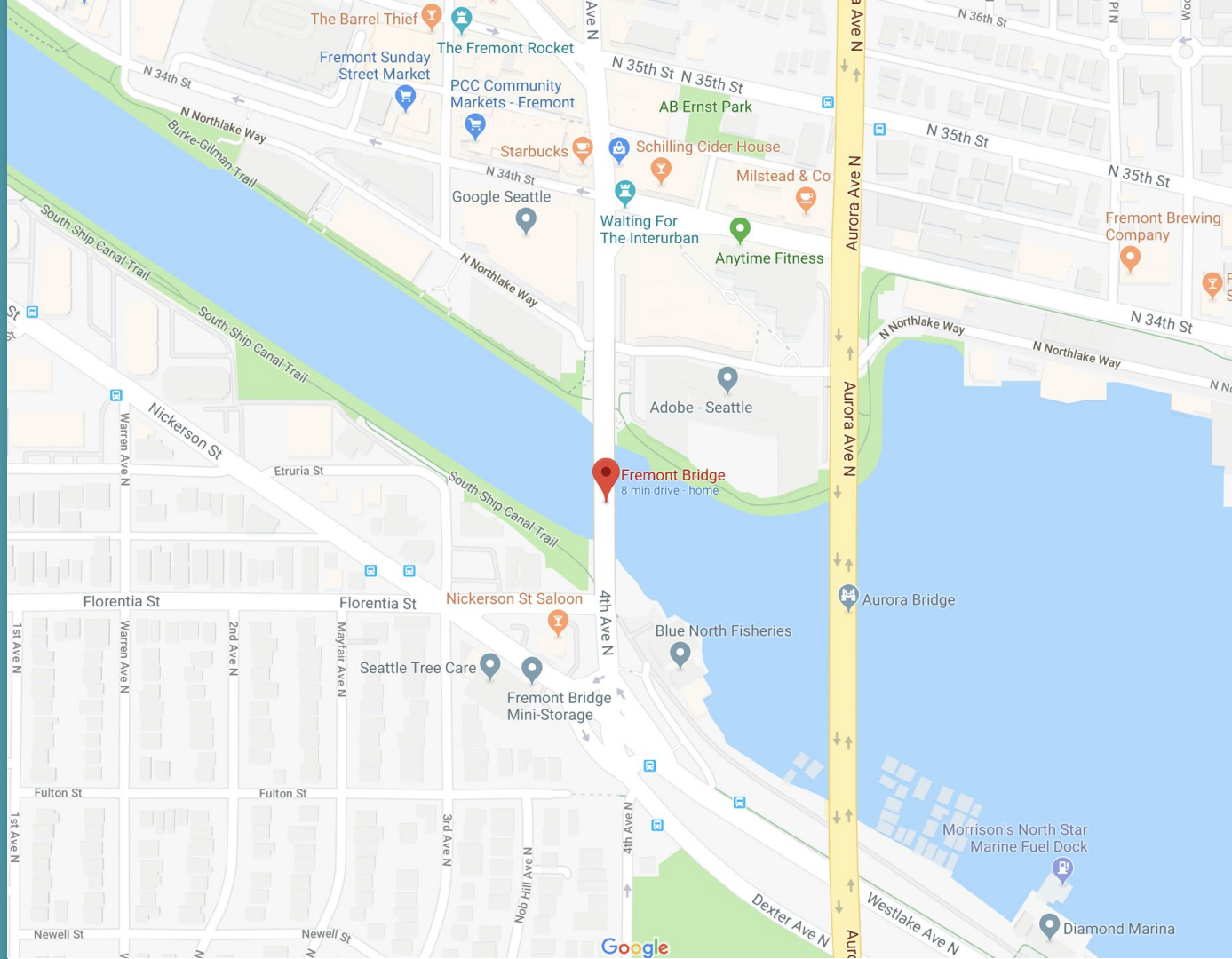
```
div(b=3, a=2)
```

Brain Break

Time Series


Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.
Exports
BALANCE in FAVOUR of ENGLAND.
Line of Imports
BALANCE AGAINST
Line of Exports
Imports
190
180
170
160
150
140
130
120
110
100,000
90
80
70
60
50
40
30
20
10
1700
1710
1720
1730
1740
1750
1760
1770
1780
The Bottom line is divided into Years, the Right hand line into £10,000 each.
Published as the Act directs, 1st May 1786. by Wm. Playfair
Neele sculpt 352, Strand, London

Fremont Bridge



Time Series

- Context: A bit more advanced than what you will need on your homework for this week, but can be helpful for your project!

- Common to change index of your data to be the timestamp
  - This allows easy querying by date

```
# Read in data with timestamp
data = pd.read_csv('data.csv', index_col='col',
                   parse_dates=True)
# Query for certain dates
data.loc['2017-03-06']   # one day
data.loc['2018-06']      # a month
data.loc['2019']         # a year
data.loc['2017':'2019']  # a range of time
```

Demo

Granularity Matters

- Your data will have a certain granularity to its time
  - e.g a row per second, a row per hour, a row per year
- Your application might require a different granularity
  - You can downsample by combining values
  - You can upsample by creating new values
  - Both are done using resample
- Use codes to describe frequency
  - D = day, W = week, M = month, A = year, ...
  - Many possible codes listed here.

- Can also groupby with time, but is not the same as downsampling.

Demo

Next Time

- Focus on data visualization
  - How to do it in Python
  - What is a “good” data viz

Before Next Time

- If you haven’t started HW2, now is a great time!